Published in final edited form as:

J Struct Biol. 2011 June ; 174(3): 425–433. doi:10.1016/j.jsb.2011.03.003.

***In vivo* virus structures: Simultaneous classification, resolution enhancement, and noise reduction in whole-cell electron tomography**

Kang Wang,

Department of Biomedical Engineering, Cornell University

Chi-yu Fu,

Department of Molecular Biology, The Scripps Research Institute

Reza Khayat,

Department of Molecular Biology, The Scripps Research Institute

Peter C. Doerschuk, and

Department of Biomedical Engineering, School of Electrical and Computer Engineering, Cornell University, 135 Weill Hall, Ithaca, NY 14853-6007, 607-255-2152

John E. Johnson

Department of Molecular Biology, The Scripps Research Institute

Kang Wang: kw243@cornell.edu; Chi-yu Fu: fuchiyu@scripps.edu; Reza Khayat: rkhayat@scripps.edu; Peter C. Doerschuk: pd83@cornell.edu; John E. Johnson: jackj@scripps.edu

Abstract

Sulfolobus Turreted Icosahedral Virus (STIV) experiences an extra-cellular environment of near boiling acid (80 C, pH 3) and particles purified under these conditions were previously analyzed by cryo electron microscopy and image reconstruction. Here we describe cryo-tomograms of *Sulfolobus* cells infected with STIV and the maximum likelihood algorithm employed to compute reconstructions of virions within the cell. Virions in four different tomograms were independently reconstructed with an average of 91 particles per tomogram and their structures compared with each other and with the higher resolution single-particle reconstruction from purified virions. The algorithm described here automatically classified and oriented two different particle types within each cell and generated reconstructions of full and empty particles. Because the particles are randomly oriented within the cell, the reconstructions do not suffer from the missing wedge of data absent from the reciprocal-space tomogram. The fact that the particles have icosahedral symmetry is used to dramatically improve the signal to noise ratio in the reconstructions. The reconstructions have approximately 60Å resolution (based on Fourier Shell Correlation analysis among reconstructions computed by the algorithm described here from four different tomograms).

© 2011 Elsevier Inc. All rights reserved.

Correspondence to: Peter C. Doerschuk, pd83@cornell.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Sulfolobus Turreted Icosahedral Virus; STIV; cryo electron tomography; cryo ET; maximum likelihood reconstruction; multiclass reconstruction; missing wedge; icosahedral symmetry

1 Introduction

Cryo-electron tomography (CET) is an imaging technique that is able, via transmission electron microscopy plus computation, to determine the 3-D structure of one-of-a-kind biological specimens up to the size of a small whole cell. When applied to a cell, CET is an *in situ* technique, i.e., it provides images of macro-molecular assemblies within the cell in a hydrated state, that is close to their native environment, and with their natural spatial relationships to other macro-molecular assemblies [16, 9]. However, CET has limitations. In order to avoid destruction of the specimen, the total electron dose is minimized, which results in a signal to noise ratio (SNR) for the original images that is typically much lower than 0.1. Furthermore, for the following two reasons, the angle coverage for the projection data is limited. First, the tilt range of the goniometer which holds the specimen is limited to about $\pm 60^\circ$ for mechanical reasons [23]. Second, the thickness of the specimen is itself a limit. Perpendicular to the sample grid, the specimen is typically $0.5\mu\text{m}$ thick. However, as the sample grid is tilted, the distance through which the electron beam must penetrate increases. For example, the penetration depth at a tilt of 60° is twice that at 0° tilt, and the penetration depth at 0° is already near the maximum achievable penetration depth. Thus, the projection images that can be acquired for CET are typically at a tilt range of $\pm 60^\circ$. By the Projection-Slice Theorem, the limited range of angles for the projection data leads to a missing wedge of data in 3-D reciprocal space which results in spatially anisotropic geometric distortion [15]. In summary, the key challenges include low SNR and limited angle coverage for the projection data which lead to a reconstructed 3-D tomogram which has low SNR and a missing wedge in reciprocal space.

In a typical infected cell there are tens to hundreds of virus particles at various stages of maturation. Therefore, if the stages can be classified and information from the particles at the same stage can be combined, both SNR and resolution can be improved. There are two fundamental challenges in combining information from different particles. First, the different particles are differently oriented and so the subcubes of the tomogram must be oriented before the information can be combined. Second, different particles come from different classes (e.g., maturation stages) and each subcube of the tomogram must first be classified before the information it contains can be combined with the information from other subcubes. Both orientation and classification are difficult due to the low SNR, the low resolution, and the missing wedge in the 3-D CET reconstruction. Particle orientation and particle classification into small numbers of homogeneous classes has been used previously in single-particle cryo-EM [5, 24, 19]. However, the data in single-particle cryo-EM and CET are quite different: The data from the CET experiment are 3-D subcubes from the tomographic reconstruction, including distortions due to problems such as the missing wedge of data, while the data in single-particle cryo-EM are 2-D projection images. These differences are the origin of two challenges. First, the 3-D nature of CET data increases the computational burden per volume relative to the computational burden per image in a single-particle computation because the number of rows in L (Eq. 6) is the number of voxels or number of pixels and the number of voxels is greater (e.g., $50^3 = 1.25 \times 10^5$ while $100^2 = 1.0 \times 10^4$). Second, the CET data is imperfect since it is itself the result of a reconstruction algorithm and most importantly has a missing wedge in reciprocal space, which hampers both the alignment and classification of particles using methods developed for single particle analysis [4].

A standard approach [5] is to classify data (images or cubes) into homogeneous classes, possibly even including orientational classes, and then reconstruct. In this type of approach, standard cross-correlation, which is used extensively in the alignment of single-particle analysis, has been shown to be not accurate in aligning data cubes with missing wedges, especially for macromolecules with high symmetry such as many viruses, which often possess icosahedral symmetry [21]. To compensate for the missing wedge, alignment algorithms based on modified cross-correlation functions that consider only the overlapping nonzero regions in reciprocal space of the two data cubes have been proposed by various researchers [1, 4, 21, 23]. Regardless of what similarity measure these algorithms use, one common feature is that the data cubes are rotated over all possible orientations in order to search for the correct alignment between each data cube and a template. Rotations will inevitably lead to interpolation of the original data cube due to the discrete nature of the data. Since the typical SNR of these tomograms is much less than 1, any form of interpolation will give rise to significant errors in the rotated data cube and hence will affect alignment accuracy. Thus, pairwise alignment of noisy data cubes may not be the optimal approach. As is described in more detail following Eq. 1, the maximum likelihood approach described here uses a mathematical model that describes the virus in continuous coordinates and rotates the model in continuous coordinates before sampling and comparing with data cubes thereby avoiding the need for interpolation.

A more fundamental issue than the issue of interpolation is that the alignment and classification depend on the accuracy of one another. Accurate alignment requires that the subset of data to be relatively homogeneous, while accurate classification requires that each data subcube be accurately aligned, so that the differences between classes are due to structural differences rather than orientational differences. One approach to the interdependency of alignment and classification is to use iterative refinement to progressively get better classifications and alignments in alternating steps (e.g., Ref. [1]). Alternatively, the alignment and classification can be performed jointly (e.g., as described in the conclusion of Ref. [4]).

In this paper, we describe an approach that goes beyond joint alignment and classification to joint alignment, classification, and restoration (SNR and resolution improvement) by maximum likelihood (ML) estimation. Joint ML classification, alignment, and reconstruction has been used extensively in single-particle cryo-EM [14, 19], but not in CET. Joint ML for CET has the same desirable properties as joint ML for single-particle cryo-EM, e.g., statistical hedging of uncertainty in class and orientation, which is important for these low SNR, low resolution 3-D data cubes. For tomograms of purified groEL/groES, a maximum likelihood approach to classification and reconstruction has been described [20] which uses a mathematical formulation in which the missing reciprocal space data is treated as nuisance parameters in an expectation maximization algorithm, which is different from the formulation in this paper, and which has not been applied to *in vivo* whole-cell tomograms.

The software that performed the calculations described in this paper is available from the authors.

2 Mathematical model for the object

The electron scattering intensity of a 3-D object, denoted by $\rho(\mathbf{x})$, can be represented as a linear combination of basis functions, denoted by $\phi_i(\mathbf{x})$, where each basis function is a continuous function of the three spatial coordinates (denoted by $\mathbf{x} \in \mathbb{R}^3$), i.e.,

$$\rho(\mathbf{x}) = \sum_i d_i \phi_i(\mathbf{x}) \quad (1)$$

where the unknown weights which determine the 3-D structure of the virus are the d_i . The selection of basis functions can be tailored specific to the macro-molecular complex of interest.

The data has discrete spatial coordinates, i.e., the indices of the voxels of tomographic cubes. However, the fact that the data is discrete does not prohibit the use of a model that describes the virus in continuous spatial coordinates. An advantage of using continuous spatial coordinates is that the model of the virus can be rotated and then sampled on a grid for comparison with the data all without any interpolation: evaluate the rotated basis function at the grid site of interest, multiply by the weight, and add. This advantage is greatest when the rotated basis function is easy to evaluate.

Sulfolobus Turreted Icosahedral Virus (STIV) is a spherical virus with icosahedral symmetry. Because the icosahedral group has 60 rotation symmetry operators and the particles are spherical, one approach to representing such an object is by a spherical harmonics series using icosahedral harmonics, but other representations of the object (e.g., voxels) are also possible. The calculations described in this paper use up to 120 basis functions (all icosahedral harmonics of order less than or equal to 25 and 10 radial basis functions for each harmonic). The choice of spherical coordinates simplifies evaluation of the rotated basis functions since the radial component of the coordinate does not change under rotations. For any representation as a linear combination of basis functions, the statistical model and ML estimation approach of Section 3 can be applied.

In this paper, the 3-D scattering intensity of a complete STIV particle is described by

$$\rho(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} h_{l,p}(|\mathbf{x}|) \Psi_{l,m}(\theta(\mathbf{x}), \varphi(\mathbf{x})), \quad (2)$$

where the unknown weights d_i of Eq 1 now require triple indices, i.e., $d_{l,m,p}$. One advantage of this model is that rotational symmetry can be built into the model by simply restricting the functions $\Psi_{l,m}$ to be a basis for the rotationally symmetric subspace of functions on the sphere, i.e., to the rotationally symmetric subspace of the space spanned by spherical harmonics. This causes a reduction in the number of parameters which need to be estimated. In the case of icosahedral symmetry, the reduction is by a factor of 60. This reduction both reduces the computational cost (Section 4) and improves the quality of the estimates. Concretely, this is achieved by defining $\Psi_{l,m}(\theta, \varphi)$ to be a linear combination of spherical harmonics $Y_{l,m'}(\theta, \varphi)$ [10, Eq. 3.53] ($m' \in \{-l, \dots, +l\}$) chosen so that the collection of $\Psi_{l,m}(\cdot, \cdot)$ span the rotationally symmetric subspace and so that $\Psi_{l,m}(\cdot, \cdot) \in \mathbb{R}$ so that $\rho(\cdot) \in \mathbb{R}$ for any choice of $d_{l,m,p} \in \mathbb{R}$. The radial basis functions, denoted by $h_{l,p}(\cdot)$ and derived by Sturm-Liouville theory [2, Ch.7], are linear combinations of spherical Bessel functions [10, Eq. 16.9] which satisfy $h_{l,p}(r) = 0$ for $0 \leq r \leq r_1$ and $r_2 \leq r$. The possibility of $r_1 = 0$ is permitted and the possibility of $r_1 = 0$ and $h_{l,p}(0) \neq 0$ for some values of l and p is also permitted. The 3-D Fourier transform of $\rho(\mathbf{x})$, denoted by $P(\mathbf{k})$, is

$$P(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \Psi_{l,m}(\theta(\mathbf{k}), \varphi(\mathbf{k})) \quad (3)$$

where $H_{l,p}(\cdot)$ is the spherical Hankel transform of $h_{l,p}(\cdot)$, which can be computed analytically because of the choice of $h_{l,p}(\cdot)$.

3 3-D cube formation model

Let $\Sigma(\mathbf{k})$ denote the 3-D Fourier transform of a noise free (i.e., ideal) subcube extracted from the CET 3-D reconstruction. Then,

$$\Sigma(\mathbf{k}) = G(\mathbf{k}) W(\mathbf{k}) \exp(-i2\pi \mathbf{k}^T \mathbf{x}_0) P(R(\alpha, \beta, \gamma)^{-1} \mathbf{k}) \quad (4)$$

where $\mathbf{x}_0$ describes the translation between the center of the coordinate system and the center of the object, (α, β, γ) are Euler angles describing the rotation of the object, $W(\cdot)$ is the 3-D frequency response describing the loss of the missing wedge of data, and $G(\cdot)$ is the 3-D CTF for the CET measurement and reconstruction process. Please note that $P(\cdot)$ depends linearly on the unknown coefficients $d_{l,m,p}$.

According to the Projection Slice Theorem, the 2-D projection of a 3-D function in a given projection direction is, after transformation to reciprocal space, a slice (i.e., 2-D plane) through the origin of the reciprocal space transformation of the 3-D function, where the plane is perpendicular to the projection direction. Thus, assuming reconstruction using a perfect interpolation filter, the missing data in the reconstruction for a single tilt experiment of $\pm\theta$ will be the region sandwiched between the two planes that form angles of $\pm\theta$ with the tilt axis. This region is shaped like a wedge, thus the term “missing wedge”. The function $W(\mathbf{k})$ in Eq. 4 accounts for the missing wedge by being 1 where the reciprocal space data is not missing and 0 where the data is missing, i.e., a binary mask. Let $\mathbf{n}_1$ and $\mathbf{n}_2$ be unit vectors normal to the two planes that are the boundaries of the missing wedge. Then

$$W(\mathbf{k}) = \begin{cases} 0, & (\mathbf{k} \cdot \mathbf{n}_1 > 1) \cap (\mathbf{k} \cdot \mathbf{n}_2 > 1) \\ 0, & (\mathbf{k} \cdot \mathbf{n}_1 < -1) \cap (\mathbf{k} \cdot \mathbf{n}_2 < -1) \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

where $\cap$ indicates the logical “and” operation.

Assuming that the virus particles are roughly randomly oriented relative to the electron beam of the microscope and are sufficiently numerous, it is expected that no region of reciprocal space will be in the missing wedge of all the particles. Therefore, by combining information from different particles, the method described in this paper can greatly reduce the effect of the missing wedge.

In the current software, the 3-D CTF for the CET measurement and reconstruction process, introduced in Eq. 4 and denoted by $G(\mathbf{k})$, is assumed to be 1 at all spatial frequencies. Therefore, the work described in this paper accounts for the low SNR, the low resolution, and the missing wedge of the CET result but not other limitations of the CET result. Such limitations, if linear, could be incorporated in $G(\mathbf{k})$.

As discussed previously, the CET cube has quite low SNR. The 3-D cube noise is assumed to be an additive white zero-mean known-variance Gaussian noise where the variance is actually estimated in a preliminary calculation. The assumption of additive Gaussian noise is motivated by the simplicity of the solution of the estimation problem formulated in Section 4.1.

Because P in Eq. 4 is a linear function of the unknown coefficients $d_{l,m,p}$, Eq. 4 coupled with the noise assumptions of the previous paragraph can be rewritten as follows. Let the pixels of a reciprocal space subcube be arrayed in a vector denoted by y and let the voxel noise be similarly arrayed in a vector denoted by v . Let the unknown coefficients $d_{l,m,p}$ be arrayed in a vector d . Let $z = (\mathbf{x}_0, \alpha, \beta, \gamma)$. Then Eq. 4 and the noise assumptions of the preceding paragraph imply that there exists a matrix, denoted by L which depends on z , such that the relationship between y , d , and v is

$$y = L(z)d + v. \quad (6)$$

The elements of L are defined by Eq. 4 and are products of icosahedral harmonics and radial basis functions. When multiple classes are considered, there are multiple d vectors, one for each class, and potentially multiple L matrices if, for example, different classes have different symmetry assumptions and therefore different dimensions for d even though they achieve the same resolution.

4 Statistical estimation

4.1 Maximum likelihood estimation

Eq. 6 implies a likelihood function for estimating d from y by maximum likelihood estimation, in particular, the likelihood function conditional on z is $p(y|d, z) = \mathcal{N}(L(z)d, \Sigma)$ where Σ is the covariance of v and $\mathcal{N}(m, V)$ is the multivariable Gaussian probability density function (pdf) with mean vector m and covariance matrix V . Then the unconditional likelihood function (which can be computed once the pdf on z , denoted by $p(z)$, is specified) is

$$p(y|d) = \int_z p(y|d, z)p(z)dz. \quad (7)$$

The pdf for z which is used in Section 5 is uniform over a 3-D sphere in $\mathbf{x}_0$ and uniform over all orientations in (α, β, γ) (i.e., Haar measure on SO_3). The subcube extracted from the tomogram is a cube to which a spherical mask is applied in order to remove objects adjacent to the virus particle. Before masking, Σ is proportional to the identity, but masking introduces correlation. However, because the mask is large, the correlation is mostly between nearest-neighbor pixels and is not accounted for in our current software.

Given the likelihood function (Eq. 7), the goal is to estimate d^η for each class where class is indexed by η . To simplify notation, only the equations for a one-class problem are written so the index η is not necessary. For a multi-class problem with N_η number of classes, each iteration of the EM algorithm involves computing the same integrals (Eqs. 13 and 14) and solving the same system of linear equations (Eq. 12) as was done in the one-class problem but now for each of N_η different classes. In the maximum-likelihood sense, the estimate of d , denoted by $\hat{d}$, is the solution that maximizes the likelihood function (Eq. 7), i.e.,

$$\hat{d} = \arg \max_d p(y|d). \quad (8)$$

Using the same type of approach as was employed in Ref. [3], an expectation maximization (EM) algorithm is derived for computing the maximum likelihood estimate for d . An EM algorithm is an iterative algorithm that progressively finds a better estimate of the model parameter d at each iteration. Specifically, at iteration n of the EM algorithm, a quantity Q is computed based on the estimate of the model parameter from the previous step, denoted by d_{n-1} , where Q is defined by

$$Q(d|d_{n-1}, y) = \int_z (d^T L^T L d - 2y^T L d) p(z|d_{n-1}, y) dz \quad (9)$$

$$= d \left\{ \int_z L^T L z p(z|d_{n-1}, y) dz \right\} d - 2 \left\{ \int_z y^T L(z) p(z|d_{n-1}, y) dz \right\} d. \quad (10)$$

The second equality is due to the linearity of the integral and the fact that d does not depend on z . Then the estimate of the model parameter at the n th step, denoted by d_n , is

$$d_n = \arg \max_d Q(d|d_{n-1}, y). \quad (11)$$

Notice that Eq. 10 is a quadratic form in d , for which the solution that maximizes the quantity Q can be obtained by solving the linear system

$$g = F d_n \quad (12)$$

where

$$F = \int_z L^T(z) L(z) p(z|d_{n-1}, y) dz \quad (13)$$

$$g = \int_z L^T(z) y p(z|d_{n-1}, y) dz. \quad (14)$$

4.2 Practical issues

Similar to most cryo EM reconstruction algorithms, the resolution of the whole-particle reconstruction is increased in a series of steps. Resolution of the model is controlled by truncating the l and p sums in Eq. 2 to $0 \leq l \leq l_{\max}$ and $1 \leq p \leq p_{\max}$. Five reconstructions of progressively improved resolution, denoted Step 0 to Step 4, are calculated, in which the $l_{\max}$ and $p_{\max}$ values for each step are the same as were used in Ref. [24].

Step 0 is estimation of a spherically symmetric model (i.e., only the $l = 0$ $d_{l,m,p}$ coefficients in Eq. 2 are used) based on the available data. In this case, no knowledge of the particle

orientation for each data cube is needed and therefore a one-class reconstruction is essentially a linear least squares problem which is solved by standard methods. Steps 1–4 use models with icosahedral but not spherical symmetry so the expectation-maximization (EM) algorithm of Section 4.1 is used. EM is an iterative algorithm that requires initialization. Since the EM algorithm converges to a local maximum of the likelihood function, multiple initializations are needed to ensure that the algorithm converges to the global maximum. One of the initial conditions is the answer from the previous step augmented with zeros for those $d_{l,m,p}$ coefficients that did not occur in the previous step. The remainder of the initial conditions are Gaussian pseudo-random perturbations of the initial condition derived from the answer from the previous step. The parameters that determine this process are identical to those in Ref. [24]. The final answer for a step is obtained by selecting the solution that has the highest likelihood value among all of the initializations.

For the whole-particle reconstructions the nuisance parameters are the class membership label of the particle and the orientation of the particle. Integration over the orientational nuisance parameters is done by the 5000 abscissa integration rule of Ref. [24].

4.3 Tests using synthetic data

As is described in Supplemental Material Section 8, the algorithm has been tested using synthetic data computed from the atomic resolution x-ray crystallographic structure of the capsid of bacteriophage Hong Kong 97 (there is no crystallographic structure of STIV). The tests include measuring performance as a function of SNR, the available tilt range, and the number of data cubes.

5 Experimental data from STIV infected *Sulfolobus sulfataricus* cells

Whole-cell tomograms of STIV infected *Sulfolobus sulfataricus* cells were collected as is described in Ref. [6]. The ML method is employed on four different tomograms in order to analyze the potential variability of assembly of viral particles in different cells and a flow chart of the calculations is given in Supplemental Material Section 7. Representative sections are shown in Figure 1. A typical viral-infected *Sulfolobus* displays multiple pyramid-like protrusions on the cell surface that cause dramatic alteration of cell morphology (Figure 1(a)). Viral particles with either full or empty cores were observed, representing DNA-filled virions and DNA-free procapsids, respectively. Virions organizing as quasi-crystalline arrays were observed, in which the array is primarily constructed of infectious particles with the procapsid particles primarily scattered outside of the arrays [6]. Some ruptured cells gushing out cytoplasmic materials were imaged (Figure 1(b–c)). On either side of the breakage sites are pyramid fragments indicating that the pyramid structure is more fragile than normal cell wall, presumably due to the absence of the hexagonally arranged surface protein layer (S-layer) in the pyramid structure. Electron-dense bodies have been observed to associate with pyramids. Apparently similar electron-dense bodies are also present in the uninfected cells, and the composition and function of these bodies remain to be determined.

A variable number of subcubes (specifically, 93, 68, 62, or 141 subcubes), each containing one virus particle, are extracted from a cell tomogram by cross-correlation with a spherically symmetric template. The missing wedge causes blurring of the outline of the virus capsid in the z direction which can be seen in x – z and y – z cross sections of the virus particles (not shown).

By manual inspection, virus particles in the cell tomogram are either packed with genome or are empty. This is not surprising because packaging kinetics can be fast (in dsDNA

bacteriophages, the time constant of the packaging kinetics is on the order of a couple minutes [7, 22, 25]), in which case only a small fraction of the particles will be in a partially-packed state. Hence, the reconstruction calculation is based on the assumption that two classes of particles exist. Furthermore, the STIV particle has icosahedral symmetry, so such a symmetry constraint is also enforced on the reconstruction.

One measure of the quality of the reconstructions is the quality of the orientations computed from the maximum likelihood estimates. One measure of the quality of the orientations is to average the oriented particles and check whether structures such as turrets are preserved in the averaging. As is shown in Figure S3c, d of Ref. [6], the turrets are preserved.

3-D visualizations of the full and empty reconstructions from two of the four tomograms are shown in Figure 2. There is a clear distinction between the reconstructions for Class 1 and 2, in particular, the Class 2 reconstruction has reduced central density and therefore represents the empty virus capsid. Thus, the algorithm is able to separate the full and empty viruses in the cell tomogram. Furthermore, the icosahedral symmetry of the capsid and the structure of the turrets at the five-fold axes of the capsid, both features that are so severely distorted in the CET cube, due to the missing wedge, that they are not visible in the x - z and y - z cross-sections of the CET cube, are clearly present in both reconstructions, providing evidence that by combining multiple noisy 3-D subcubes with different orientations, the missing data in reciprocal space is successfully recovered.

Each of the four tomograms was processed separately. Therefore there are four independent reconstructions of the full and empty particles. In order to demonstrate that the maximum likelihood estimator chooses the same class definitions in each tomogram, the Fourier Shell Correlation (FSC) was computed for all pairs of full and for all pairs of empty reconstructions with the results shown in Figure 3(a). The plots in Figure 3(a) show the FSC rising at higher frequencies after an initial drop at middle frequencies. This rising is due to the fact that there is very little energy at higher frequencies. For instance, Figure 3(b) shows a plot of

$$E(k) = \sum_{i=1}^2 \int |P_i(\mathbf{k})|^2 d\Omega', \quad (15)$$

where P_i is the reciprocal space reconstruction of the i th class and $\int d\Omega'$ is integration over angles in reciprocal space, which decreases monotonically by 6 orders of magnitude over the frequency range shown in Figure 3(a). Using an FSC cutoff of 0.5, the full (empty) particle reconstructions agree to 6.61 nm (6.67 nm) while the resolutions of the individual reconstructions (computed by FSC between even and odd numbered data subcubes) are 6.1 nm (6.9 nm). Therefore the maximum likelihood estimator, which determines its own classes once the number of classes has been set by the user, is determining the same classes for each tomogram.

All of the calculations described in this section to this point have two classes. Limited calculations using three classes have also been performed. In the three-class results, two of the resulting reconstructions are quite similar. For instance, the three spherically-averaged radial electron scattering intensity plots shown in Figure 4 are quite similar. For this reason, the other calculations reported in this manuscript are two-class calculations.

In order to better compare with established methods, e.g., Ref. [4], two additional calculations have been performed. In the first calculation, one of the subcubes is randomly

chosen as the reference subcube and the orientation of all other subcubes is determined relative to the reference subcube. The method is to maximize the constrained normalized cross-correlation [4]. The quality of the resulting orientations is qualitatively measured by using Chimera to visualize a cube containing the oriented average of the subcubes. By this measure, this process fails since the turrets of STIV are not seen in the oriented-average cube. In the second calculation, the reference subcube is a low-pass filtered version of the single-particle cryo-EM STIV structure [11]. By the same measure, this process succeeds, as is demonstrated by the oriented average shown in Figure 5. These two calculations indicate the importance of having a high-quality reference structure, which is a challenge if only CET data is available.

It is more difficult to compare with standard methods in the case of multi-class calculations because the classifier must be designed. The *k*-means classifier as implemented in Matlab¹ was used on two feature vectors derived from the 68 particles in one tomogram. The first feature vector is the $50 \times 50 \times 50$ sub-cube of voxels treated as a vector after the subcube has been oriented using the single-particle cryo-EM reference template and rotated to a standard orientation. For the first definition, the metric for comparing two vectors is the constrained normalized cross-correlation [4]. The second feature vector is the spherical average of the $50 \times 50 \times 50$ subcube of voxels. For the second definition, the metric for comparing two vectors is the Euclidean norm of the difference of the two vectors. Because there is a spherical average, it is not necessary to orient the subcube. The first vector is general in the sense that it should be sensitive to any alteration in the virus structure but it requires a template that is usually not available in order to compute the orientation. The second vector is fairly specific to alterations in the virus structure that are a function of radius only. Figure 6 shows the electron scattering intensity spherically averaged and averaged over the members of the class for both classes and both feature vectors and Table 1 compares the resulting labels with the labels from the maximum likelihood classifier. The *k*-means labels agree more closely with the maximum likelihood labels when the feature vector is designed to be sensitive to the expected distinction of empty versus full and, correspondingly, the electron scattering intensity plots show greater differences between these two classes for that feature vector.

In the STIV infected *Sulfolobus sulfataricus* cells described in Ref. [6] and in the sectional images of Figure 1, the particles tend to cluster within the cell and the clusters tend to approximately have the structure of a lattice. In Supplemental Material Section 9, algorithms are described for determining the lattice constants and the orientational heterogeneity of the particles occupying the lattice.

6 Discussion and conclusion

The Cryo-Electron Tomography (CET) studies of STIV infection in intact cells that are described in this paper reveal valuable insights into the assembly and maturation of STIV and other inner-membrane containing viruses including the fine details of ultra-structural alterations of host and virus during the infection process. With computational analysis of the cell tomogram, structures like the transiently populated procapsid form of the virus and unstable intermediates like the quasi-crystalline array of virus particles and the pyramid-like cellular structures can be visualized in their original context at low-nm resolution without the need for purification and without the concerns regarding artifacts in the specimen that go along with purification processes.

¹<http://www.mathworks.com/help/toolbox/stats/kmeans.html>.

Whole-cell CET is a unique imaging technique, in that it allows the visualization of macromolecules in their native cellular environment. However, due to technical limitations, it inherently has missing data in reciprocal space, and the missing data implies that conventional techniques do not reliably process the resulting 3-D real-space cubes. The algorithm proposed in this paper for processing CET tomograms reflects a different approach than the algorithms in Refs. [1, 4, 21], in that no explicit alignment and classification are performed on objects in the data set. Instead, a generative statistical model is used to describe each class of particles in the CET tomogram and parameters in the model are estimated from the CET tomogram by a maximum likelihood estimator. Conditional on the type of particle shown in a subcube and the orientation of the particle in the subcube, the likelihood function implied by the statistical model is essentially the dissimilar function used in Ref. [1] and, with normalization, is also the constrained cross-correlation used in Ref. [4], all of which explicitly account for the missing wedge of the data. Thus, the proposed approach shares characteristics with other approaches, but the method by which the label for the type of particle and the orientation of the particle are treated is distinct from the methods used by other approaches as is described in the following paragraph.

The proposed algorithm uses a soft alignment and classification strategy rather than the hard alignment and classification strategies used elsewhere, in that each class average is computed by averaging a weighted sum of all data cubes at all possible orientations in each step of the maximization process. The weight is proportional to the similarity of the rotated data cube and the class average from the previous iteration. As the class average improves, the averaging will be weight mostly toward the data cubes within that class and the correct orientation of each data cube. Thus, as the algorithm converges, the underlying structures will be obtained. This alignment and classification strategy might be most important at the initial stage, when the class averages are poor. In this situation an algorithm can lock onto grouping data cubes that contain particles with similar missing wedge orientations rather than similar structures. However, with the soft classification, the class average might not have a strong bias in this situation, making it easier to escape such a local minimum.

The maximum likelihood approach is robust against noise, in particular, in simulation, it performs relatively well even at an signal to noise ratio (SNR) of 0.005. As the whole-cell CET technology matures, better cameras will allow the acquisition of projection images at finer pixel sampling rates which means that the number of electrons per pixel will be reduced in order to keep the total electron dose from growing. This will result in CET tomograms with smaller voxels, higher resolution, and lower SNR. Thus, as the field evolves toward higher resolution, it is crucial for CET tomogram processing algorithms to have excellent performance at low SNR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Bartesaghi A, Sprechmann P, Liu J, Randall G, Sapiro G, Subramaniam S. Classification and 3D averaging with missing wedge correction in biological electron tomography. *Journal of Structural Biology*. June.2008 162:436–450. [PubMed: 18440828]
2. Coddington, EA.; Levinson, N. *Theory of Ordinary Differential Equations*. McGraw-Hill; New York: 1955.
3. Doerschuk PC, Johnson JE. *Ab initio* reconstruction and experimental design for cryo electron microscopy. *IEEE Trans Info Theory*. Aug; 2000 46(5):1714–1729.10.1109/18.857786

4. Forster F, Pruggnaller S, Seybert A, Frangakis AS. Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of Structural Biology*. 2008; 161:276–286. [PubMed: 17720536]
5. Frank, J. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press; San Diego: 1996.
6. Fu, Cy; Wang, K.; Lanman, J.; Khayat, R.; Young, MJ.; Jensen, GJ.; Doerschuk, PC.; Johnson, JE. In vivo assembly of an archaeal virus studied with whole cell electron cryotomography. *Structure*. December 8.2010 18:1579–1586. [PubMed: 21134637]
7. Fuller D, Raymer D, Kottadiel V, Rao V, Smith D. Single phage t4 dna packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proc Nat Acad Sci USA*. 2007; 104:16868–16873. [PubMed: 17942694]
8. Gan L, Speir JA, Conway JF, Lander G, Cheng N, Firek BA, Hendrix RW, Duda RL, Liljas L, Johnson JE. Capsid conformational sampling in HK97 maturation visualized by x-ray crystallography and cryo-EM. *Structure*. Nov; 2006 14(11):1655–1665. [PubMed: 17098191]
9. Grunewald K, Desai P, Winkler DC, Heymann JB, Belnap DM, Baumeister W, Steven AC. Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*. 2003; 302:1396–1398. [PubMed: 14631040]
10. Jackson, JD. *Classical Electrodynamics*. 2. John Wiley; New York: 1975.
11. Khayat R, Tang L, Larson ET, Lawrence CM, Young M, Johnson JE. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proc Nat Acad Sci USA*. 2005; 102(52):18944–18949. [PubMed: 16357204]
12. Kremer JR, Mastronarde DN, McIntosh JR. Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology*. 1996; 116:71–76. [PubMed: 8742726]
13. Lancaster, P.; Tismenetsky, M. *The Theory of Matrices*. 2. Academic Press; 1985.
14. Lee J, Yin Z, Doerschuk PC, Tang J, Johnson JE. Automatic *ab initio* simultaneous classification and 3-D reconstruction of multiple types of viruses from cryo electron microscope images showing a mixture of all types. Submitted to *J. Struct. Biol*.
15. Leis AP, Beck M, Gruska M, Best C, Hegerl R, Wolfgang B, Leis JW. Cryo-electron tomography of biological specimens. *IEEE Sig Proc Mag*. 2006; 23:95–103.
16. Lucic V, Forster F, Baumeister W. Structural studies by electron tomography: From cells to molecules. *Annual Review of Biochemistry*. 2005; 74:833–865.
17. Mastronarde DN. Dual-axis tomography: an approach with alignment methods that preserve resolution. *Journal of Structural Biology*. 1997; 120:343–352. [PubMed: 9441937]
18. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem*. 2004; 25(13):1605–1612. [PubMed: 15264254]
19. Scheres SHW, Gao H, Valle M, Herman GT, Eggermont PPB, Frank J, Carazo JM. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*. Jan; 2007 4(1):27–29. [PubMed: 17179934]
20. Scheres SHW, Melero R, Valle M, Carazo J-M. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure*. December 9.2009 17:1563–1572. [PubMed: 20004160]
21. Schmid MF, Booth CR. Methods for aligning and for averaging 3D volumes with missing data. *Journal of Structural Biology*. 2008; 161:243–248. [PubMed: 18299206]
22. Smith D, Tans S, Smith S, Grimes S, Anderson D, Bustamante C. The bacteriophage straight phi29 portal motor can package dna against a large internal force. *Nature*. 2001; 413:748–752. [PubMed: 11607035]
23. Winkler H, Zhu P, Liu J, Ye F, Roux KH, Taylor KA. Tomographic subvolume alignment and subvolume classification applied to myosin V and SIV envelope spikes. *Journal of Structural Biology*. 2009; 165:64–77. [PubMed: 19032983]
24. Yin Z, Zheng Y, Doerschuk PC, Natarajan P, Johnson JE. A statistical approach to computer processing of cryo electron microscope images: Virion classification and 3-D reconstruction. *J Struct Biol*. 2003; 144(1/2):24–50. [PubMed: 14643207]

25. Ziedaite G, Kivela H, Bamford J, Bamford D. Purified membrane-containing procapsids of bacteriophage ϕ 230 package the viral genome. *J Mol Biol.* 2009; 386:637–647. [PubMed: 19150363]

Figure 1.

Slices from a whole cell ECT of *S. solfataricus* infected with STIV. Panel (a) A 10 nm slice of a 3-D tomogram in which the slice is perpendicular to the direction of the beam. Multiple pyramid-like protrusions are present. Both DNA-free procapsid and DNA-filled virions are present in the cytoplasm. A quasi-crystalline viral array is visible in the upper right quadrant of the cell. Scale bar, 200 nm. Panels (b) and (c) Enlarged views of breakage sites in the ruptured cell. The fragments reminiscent of pyramid structures, i.e., fragments of cell wall which lack the S-layer of typical cell wall, are visible adjacent to the rupture site.

Figure 2. Reconstructions, visualized by Chimera [18], of full and empty particles from two of the four tomograms. Panels (a) and (b): First tomogram. Panels (c) and (d): Second tomogram.

Figure 3.

Resolution of the reconstructions. Panel (a): Fourier Shell Correlation (FSC) curves between all pairs of full-particle reconstructions and empty-particle reconstructions. The high frequency cutoff of the plots is the Nyquist frequency (the voxel size is 1.24nm so the Nyquist frequency is $1/(2 \times 1.24)\text{nm}^{-1} \approx 0.4\text{nm}^{-1}$). Panel (b): Base-10 logarithm of the class and angular average of the reciprocal space energy ($\log_{10} E(k)$ of Eq. 15) as a function of the magnitude of the spatial frequency vector. The low energy of the reconstruction at large spatial frequencies leads to spurious fluctuations in the FSC curves of Panel (a).

Figure 4. Spherically-averaged radial electron scattering intensity curves for the three reconstructions of STIV resulting from a three-class calculation. The horizontal axis is fractional distance to the outer radius of the particle.

Figure 5.

The structure resulting from alignment of each subcube using a reference template from single-particle cryo EM followed by rotation to a standard orientation and averaging. Note that the turrets are present, indicating that the orientations are accurate. This indicates the importance of using a high-quality reference template.

Figure 6. Averages over the sphere and over the members of the class determined by the k -means labels for two different feature vectors. Red: full particles. Blue: empty particles. The horizontal axis is fractional distance to the outer radius of the particle.

Table 1

Comparison of the labels from the k -means classifier and the maximum likelihood classifier.

(a) feature vector: voxels of the subcube			
		k-means labels	
		empty	full
ML labels	empty	17	20
	full	12	19

(b) feature vector: spherical average of the voxels of the subcube			
		k-means labels	
		empty	full
ML labels	empty	20	10
	full	10	28